Holden Bridge
Professor Basit
CS 5010
July 23, 2020

HW3 Python and Web Scraper Report

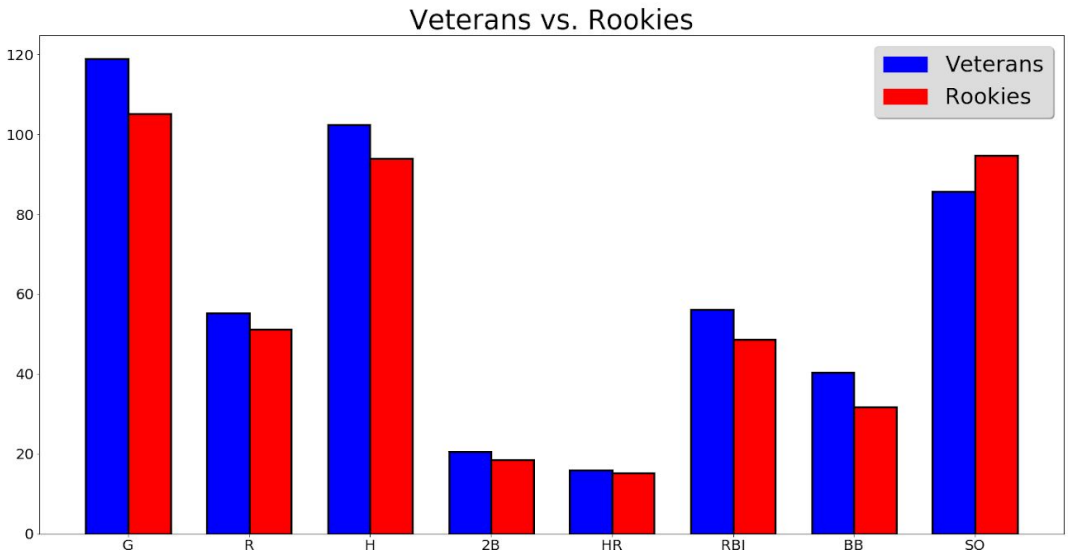
In this assignment I wrote a Python script, also known as a web scraper, that extracted Major League Baseball (MLB) hitting statistics from ESPN and saved it as a dataframe using Python's pandas library. Then, I turned the MLB data I collected into workable "knowledge" by giving the data a story. My approach to this assignment was to divide it into three easy to manage parts. Part 1 was writing the web scraper and saving the data into a pandas dataframe. Part 2 was analyzing the data and deciding what story I wanted to tell with it. Part 3 was the drafting of this report to tie everything together.

The data I scraped was offensive statistics from MLB every player in the 2019 season. Each row in the dataframe consisted of a player's name (qualitative data) and 14 numerical baseball statistics produced by that player (quantitative data). I labeled each code block of my jupyter notebook file and will use that labeling to refer to specific portions of my code:

- Code Block 1: In this block, I imported the python libraries I used throughout the assignment. This included BeautifulSoup for the web scraping, pandas for the dataframe, and matplotlib and pyplot for data visualization as well as others.
- Code Block 2: In this block, I used BeautifulSoup to read the webpage's HTML code. I then wrote a couple lines to practice using the .find() and .find_all() functions for parsing the HTML and finding the fields I wanted to save before tackling gathering all of the data.
- Code Block 3: In this block, I first created a pandas dataframe and pulled the correct column headers from the webpage. I then looped through all 346 players and saved their stats into the dataframe. There are some comments about the specifics of the loop in the block.

The question I set out to answer was, do veterans (players with 10+ years in the league) or rookies (players in their first or second year) perform noticeably better than the other group?

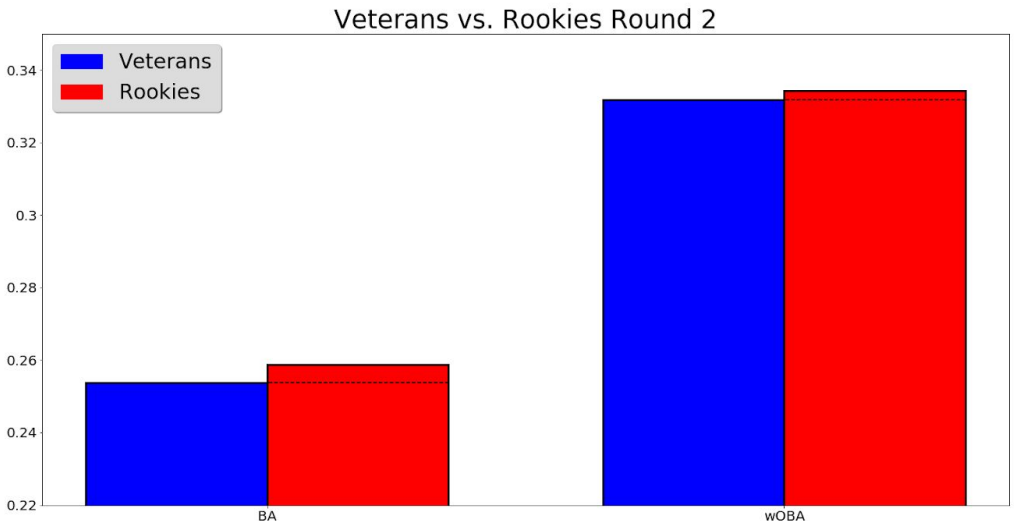
- Code Block 4: In this block I converted the statistics in the dataframe to the correct types because they were initially formatted as strings rather than ints or floats. Then, I created the two subframes for veterans and rookies. Next, I created a grouped bar plot to visualize the results.



Looking at the plot it appears that the MLB veterans outperformed the rookies in all of the displayed statistics. Given my knowledge of baseball, I was slightly surprised by these results, so I wanted to dig a little deeper into the data. If you look at the far left column you will see that veterans on average played 10-15 more games than rookies did during the season. Could this explain why veterans appeared to perform better than the rookies? I have worked with baseball data before and know that traditional statistics are great for summarizing a player's season but are actually not the best indicator of performance. I decided to add one of my favorite statistics, wOBA (weighted on base average) to the dataframe. Simply put, wOBA weights the outcome of each at bat so that players are rewarded more for doubles and home runs and combines this into one handy number to represent a player’s total offensive value.

- Code Block 5: Here I add 1B (singles) to the dataframe because it is needed for the wOBA calculation as well as the wOBA statistic itself.

- Code Block 6: In the last code block, I used the same code I used to create the first grouped bar plot but only for batting average and wOBA this time. I added dashed lines to help illustrate the difference between the rookies and veterans.



Looking at this plot it is likely that rookies actually performed just as well if not slightly better than the veterans. This is an excellent example of diving deeper into the data and not trusting the first result we get.

Somebody could use my program to pull MLB statistics and manipulate them however they would like. I only touched on a small subset of the data so there are many more discoveries that can be made. This program can be improved by pulling stats from a more robust source. The ESPN link only has 14 stats for each player which was sufficient for this assignment, but many more stats are calculated and tracked.

Sources:

ESPN Link - http://www.espn.com/mlb/history/leaders/_/breakdown/season/year/2019/start/1

wOBA Source - https://library.fangraphs.com/offense/woba/